The Case For Anomalous Link Discovery

Matthew J. Rattigan
University of Massachusetts
140 Governors Drive
Amherst, MA 01003 USA
rattigan@cs.umass.edu

David Jensen
University of Massachusetts
140 Governors Drive
Amherst, MA 01003 USA
jensen@cs.umass.edu

ABSTRACT

In this paper, we describe the challenges inherent to the task of *link prediction*, and we analyze one reason why many link prediction models perform poorly. Specifically, we demonstrate the effects of the extremely large class skew associated with the link prediction task. We then present an alternate task — *anomalous link discovery* (ALD) — and qualitatively demonstrate the effectiveness of simple link prediction models for the ALD task. We show that even the simplistic structural models that perform poorly on link prediction can perform quite well at the ALD task.

Keywords

Link prediction, anomalous link discovery, relational learning.

1. INTRODUCTION

Modeling the link structure of relational data is a key challenge of relational learning. To date, many link modeling efforts have focused on the task of *link prediction*: given a dynamic graph representing objects and their relationships, predict which new relationships will appear in the near future [7][10]. A fundamental challenge in link prediction is a highly skewed class distribution — as networks grow and evolve, the number of negative examples (disconnected pairs of objects) increases quadratically while the number of positive examples often grows only linearly. Thus, the evaluation of a link prediction model is hampered by the computational cost of evaluating all possible pairs of objects for which a link might exist, and the highly skewed class distribution increases the variance of the model. As a result, most attempts at accurate link prediction have produced models with very low accuracy.

We propose the alternative task of *anomalous link discovery* (ALD): given a static or dynamic graph representing objects and their relationships, identify those links that are anomalous. In this paper, we equate anomalous links with those that are statistically unlikely, and we contend that, in many cases, the most "interesting" links in the data are those that are statistically unlikely.

A myriad of link models can be found in the literature [4][7][9][10][11]. They vary in complexity and effectiveness at probabilistically modeling link structure in relational data. Many such models, especially those that rely on link structure alone, can be crippled by the variance introduced by imbalanced data. However, as we will see below, even a very simple model such as one based on the Katz measure [7] can be very effective at the ALD task.

1.1 Link Prediction

Probabilistic modeling of link structure is a growing area of research in statistical relational learning [3]. Many algorithms for attacking this problem approach it as a binary classification task in which pairs of objects in the data comprise instances. Given any pair of objects, the model assesses the probability of a new link appearing between them at some point in the future. While some link models found in the literature take into account object attribute information [11], the models we consider in this paper use only graph structure, because such models are sufficient to examine the effects we wish to investigate.

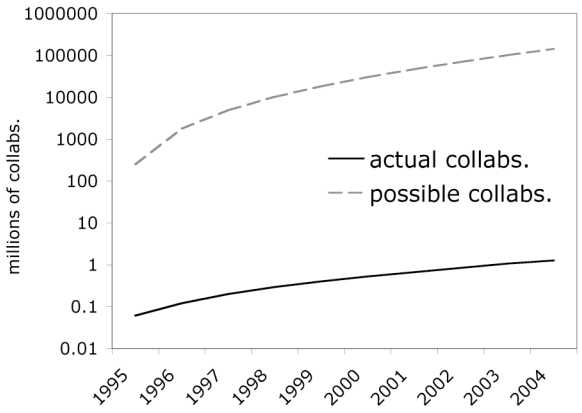


Figure 1. Logarithmic plot of actual and possible collaborations between DBLP authors, 1995-2004.

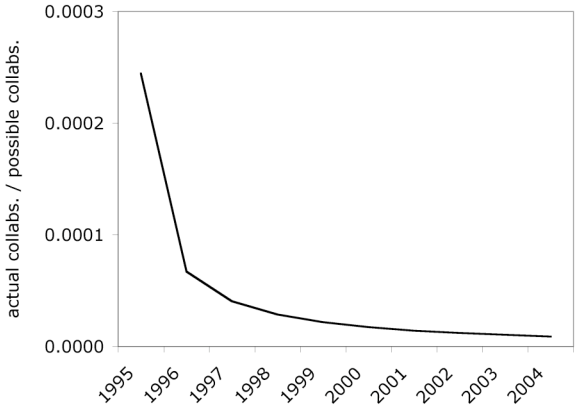


Figure 2. Publications of DBLP authors as a proportion of possible collaborations, 1995-2004.

While this approach to link modeling seems straightforward, the combinatorics of the task poses formidable challenges. The number of possible links is quadratic in the number of objects, but in many domains where link prediction is employed the number of actual links added to the graph in any time period is only a tiny

fraction of this number. The result is that algorithms for link modeling must contend with a extremely large class skew, making both learning and inference difficult [5]. Figure 1 illustrates the problem for data from the Digital Bibliography & Library Project (DBLP) over a ten-year span. During this time period, the number of authors increased by a factor of 13, from approximately 22 thousand to 286 thousand, and the number of possible collaborations increased by a factor of 169. However, the number of actual collaborative papers increased by only a factor of 21.

Figure 2 depicts the proportion of possible collaborations between authors that occur during this same time period. This ratio represents the class distribution for the pairwise link prediction problem. By 2004, less than one-thousandth of one percent of DBLP author pairs had written a paper together.

In the face of such an uneven class distribution, even the very accurate binary classifiers can have extremely low raw accuracy in predicting link structure [7]. The high error rate results from a combination of the variance in the model's estimates and the imbalance in the class distribution. The high proportion of negative instances (unlinked object pairs) means that even when a relatively low percentage of negative instances are assigned scores similar to the positive instances, the model will produce an extremely large raw number of false positive inferences.

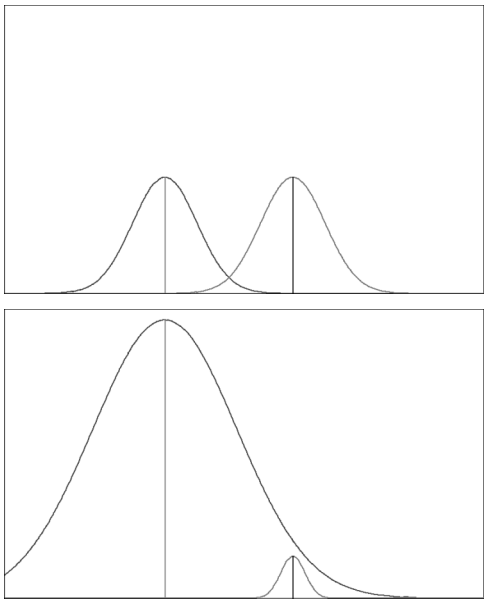


Figure 3. (top) Schematic representation of the effects of large class skew on a model's ability to discriminate between classes. In the first case, the two distributions are easily distinguished. (bottom) In the second case, large class skew virtually eliminates our ability to produce highly accurate predictions.

We illustrate the problem schematically in Figure 3. Consider a hypothetical data set and some statistic *s* that is measured on each instance pair. We assume that the values of *s* are drawn from separate distributions for linked (positive) and non-linked (negative) object pairs, illustrated in Figure 3 (top) as Normal distributions with differing means. In the face of large class skew, the entirety of the positive class distribution is "swallowed" by the tail of the negative class distribution, as seen in Figure 3 (bottom).

In the latter case, it is virtually impossible produce highly accurate predictions by examining values of *s*. Attempts to overcome the variance issues will usually introduce additional bias to the classifier [2].

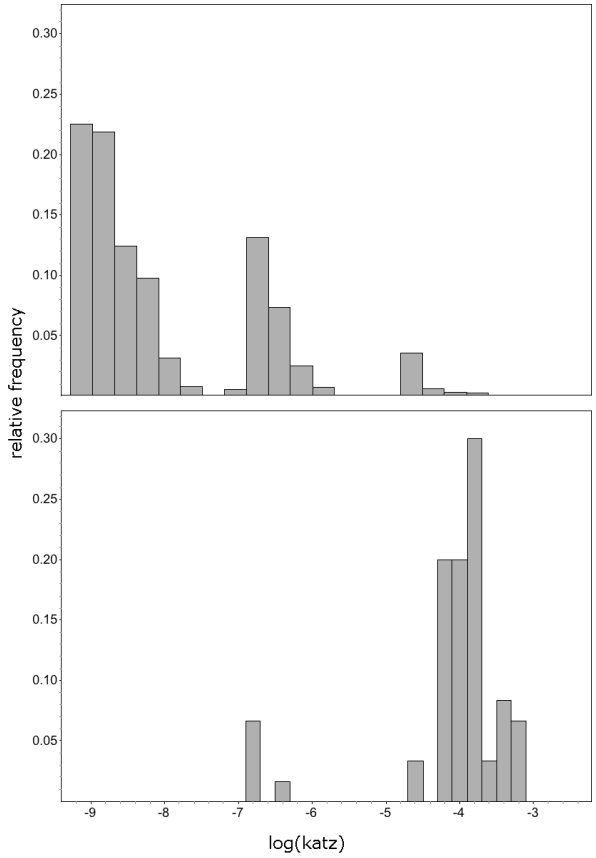


Figure 4. Relative frequency distributions of the log-Katz statistic for author pairs that include Jiawei Han for all possible coauthors (top) and actual coauthors (bottom).

As an example, consider the case of Jiawei Han, a professor at the University of Illinois at Urbana-Champaign and a KDD 2005 program committee member. We took a snapshot of the DBLP database as it existed in 2002, and identified 39 thousand "core authors" (authors who had several publications in the eight preceding years) much in the same manner as Liben-Nowell and Kleinberg[7]. In the standard link prediction task, our goal is to predict which authors Prof. Han will collaborate with in the future (e.g., 2003-2004). To discriminate between positive coauthor pairs (those that Prof. Han does write a paper with) and negative pairs, we use the Katz measure (see Section 2), which Liben-Nowell and Kleinberg have shown to be a (relatively) effective statistic for link prediction tasks [7]. The relative frequency distributions for each class can be seen in Figure 4. Clearly, the values of the Katz measure for Prof. Han's actual coauthors are drawn from a different distribution than the author population at large. This would lead us to believe that an accurate model for

link prediction can be constructed fairly easily.However, the absolute frequency distributions shown in Figure 5 tell a different story, showing the effects of the large class skew. The distributions depicted are the same as those shown in Figure 4, but plotted with actual frequency counts rather than relative ones. According to DBLP, Prof. Han collaborated with 63 different core authors in 2003-2004 (out of a possible 39,000), and it becomes virtually impossible to differentiate scores for positive instances from the tail of the score distribution for negative instances (the problem is actually even worse than depicted, as computational constraints allowed us to calculate the Katz measure for only 10,000 possible coauthors). We see here how the variance associated with the class imbalance can drown out the separability achieved by the simple model. Thus a Katz-based ranking classifier that draws a classification threshold generous enough to capture even half of the actual positives will misclassify hundreds of negative pairs.

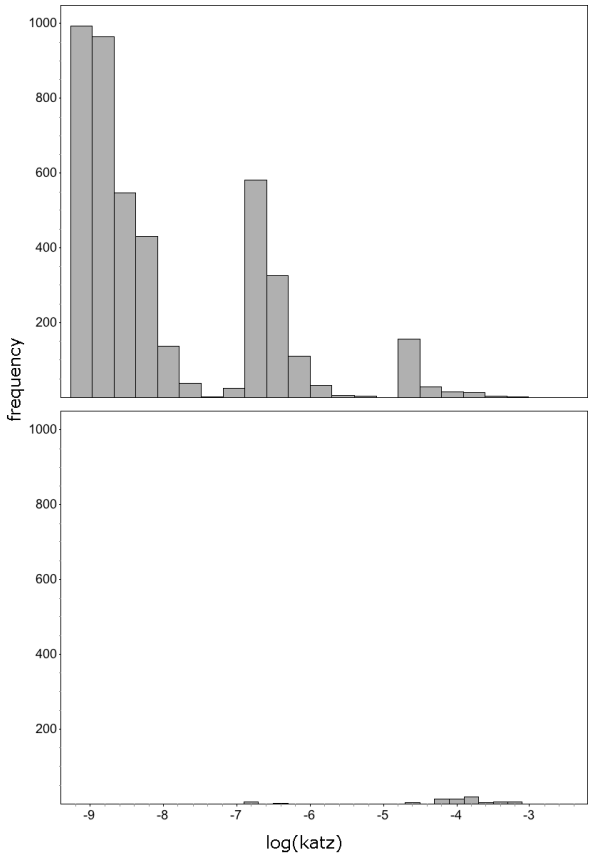


Figure 5. Actual frequency distributions of the log-Katz statistic for negative Jiawei Han author pairs (top) and positive author pairs (bottom).

The results of these effects are models of questionable utility. The vast majority of new links are not correctly predicted, and the number of false negatives is quite substantial [7]. In addition, the links that are predicted correctly tend to be the "obvious" ones, those that are out of the "reach" of the tail of the massive negative sampling distribution.

Finally, this sort of pairwise link prediction is often computationally intractable [9]. Given new objects in the graph, it may not be possible to efficiently consider their link probabilities with all other objects. Again, attempts to restrict the search introduce bias that may eliminate the possibility of predicting the most "interesting" links.

Given the issues described above, the link prediction task is a daunting undertaking. Furthermore, due to the challenging combinatorics, it may very well be impossible to satisfactorily address link prediction as currently specified. In the face of these facts, what is the aspiring link modeler to do? One potential avenue is to refocus structure learning efforts away from predicting link existence and toward predicting properties of existing links. Specifically, we propose identifying anomalous links that actually appear in the data. By respecifying the task, we can leverage existing algorithms to gain insight into the structure of graphs.

1.2 Anomalous Link Discovery

In some ways, anomaly discovery can be seen as the most abstract of knowledge discovery tasks. The goal is to identify "outliers" in a data set. In traditional data mining, anomaly detection algorithms are most often employed in domains that deal with security issues, picking out suspicious login sessions at a computer or strange financial transactions.

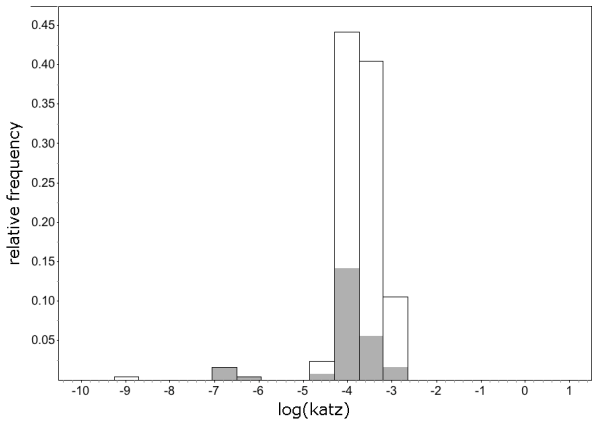


Figure 6. Relative frequency distribution of the log-Katz measure for Jiawei Han's collaborations, 1995-2004. The shaded region represent the collaborations from the final two years.

In the case of relational learning, anomaly detection involves modeling the likelihood of links that appear in the data. Relational learning techniques seem especially suited to the anomaly detection problem, because structured data lend themselves to a host of possible methods for finding interesting instances in a data set. As we have seen, the interesting facets of our data may be defined by the relations between entities as well as the intrinsic properties of the entities themselves. Despite this seemingly natural alignment of tasks and techniques, anomaly detection tasks have been all but ignored in the literature on relational learning [8].

Furthermore, in many domains, identifying anomalous links may actually be more useful than predicting links. For instance, bibliometric data are often the target of link prediction[3]. However,

is it more useful to "predict" coauthorships and citation, or to realize when an interesting (or unlikely) collaboration occurs? In the case of Internet applications, is predicting which web pages will potentially link to each other very helpful? Or would we rather be able to comb the new links that are created each day and be alerted when surprising connections are made? Obviously, the utility of ALD algorithms is domain dependent, but even a cursory examination of the literature suggests several possible applications. Finally, ALD models scale well with the size of the graph, unaffected by an ever-widening class distribution gap.

Returning to the example of Prof. Han, we can examine the properties of the collaboration links that appear in 2003-2004, and examine their likelihood given past collaborations. Figure 6 shows the distribution of the Katz measure for all of Prof. Han's collaborations with core authors dating back to 1995. Using this distribution, we can rank the new collaborations by likelihood.

On one extreme, we have a collaboration between Jiawei Han and Jian Pei of SUNY Buffalo, which was measured to be the most likely collaboration involving Prof. Han with a Katz score of 6.96×10^{-4}. An examination of the authors' home pages bears this out --- the two have written dozens of papers together over the years, thus their recent joint work should not surprise anyone.

However, not all of Dr. Han's collaborations are so predictable. During the same year, he authored a paper with Martin Ester of Simon Fraser University. According to a the Katz measure, this is one of the more unlikely of Prof. Han's collaborations in 2003-2004, with a score of 4.3×10^{-7}. Both authors are accomplished scientists, yet until last year they were only connected in the authorship graph by paths of four or more hops. A visual representation of the subgraphs connecting the two pairs of authors is shown in Figure 8.

2. PRELIMINARY RESULTS

To test the validity of our intuitions concerning the use of link prediction techniques for anomaly detection, we examined the author graph drawn from DBLP data. In our representation, we have a single object type, representing authors. Journal and conference papers are represented as links that connect the authors together through coauthorships. Thus a single paper with four authors will spawn ten different items in the graph: a node for each author, and six links expressing the pairwise coauthorships between them. A visual representation of the relational schema can be seen in Figure 7. Note that for the purposes of our experiments, our models operate on the link structure alone, and do not incorporate any paper attribute information.

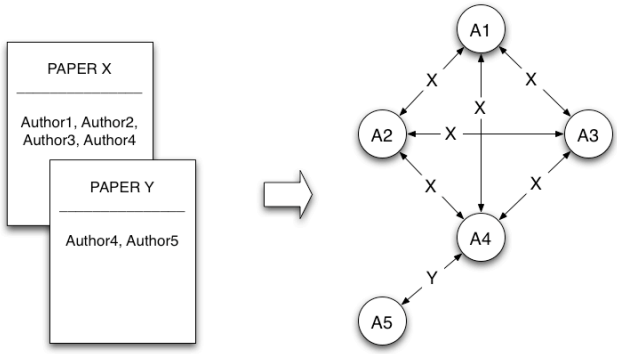


Figure 7. Relational schema DBLP data: authors are represented by author that are linked through co-authored papers.

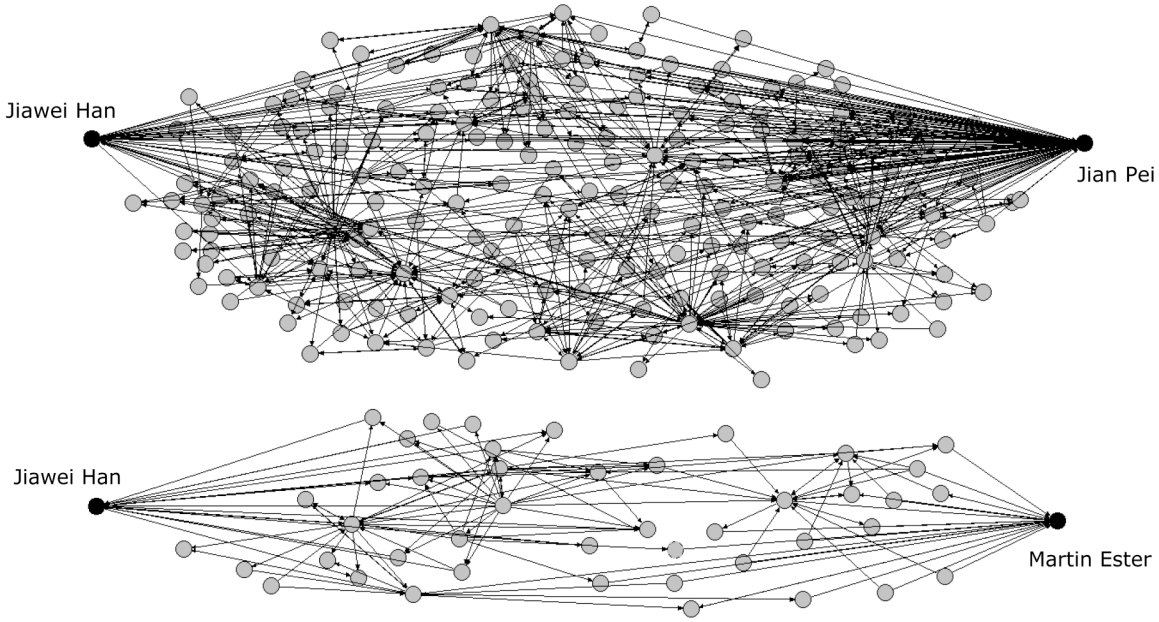
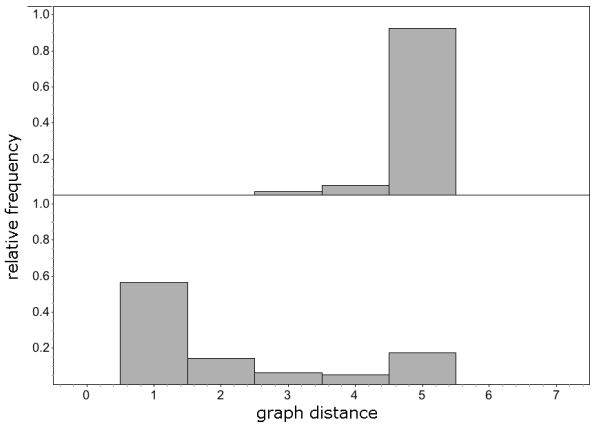
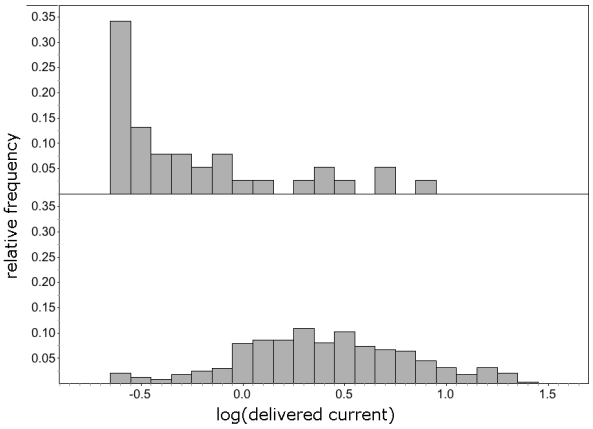


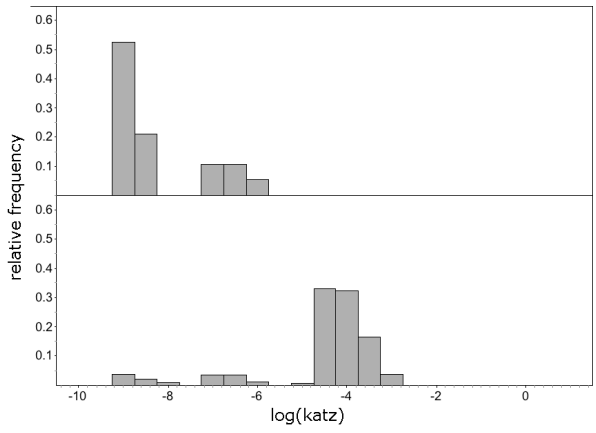
Figure 8. Subgraphs made up of all paths of length four or less connecting Jiawei Han with Jian Pei (top) and Martin Ester (bottom). The Katz scores for the two subgraphs differ by three orders of magnitude.



(a)



(b)



(c)

Figure 9. Sampling distributions for linked (bottom) and unlinked (top) author pairs for three measures of link likelihood: (a) graph distance, (b) delivered current, and (c) Katz score.

We tested several measures identified in the link prediction literature as being useful at predicting links [7]. The histograms in Figure 8 illustrate the "discriminative ability" for three of these measures. In each case, the sampling distribution for negative (unlinked) pairs is shown directly above the distribution of pairs that contain links.

The first statistic shown is simple graph distance between the objects in the pair. While the linked and unlinked values are clearly drawn from different distributions, there is still a great deal of overlap between the two.

The second measure depicted, called *delivered current*, is a measure of the maximum deliverable current when the nodes in the subgraph connecting the start and end nodes are treated as wired resistors in an electric circuit[1]. A quick comparison shows this measure to be more useful than simple graph distance.

Finally, we show the positive and negative distributions for the Katz measure[7] used in the examples in section 1.2. The Katz measure is a weighted sum of the number of paths in the graph that connect two nodes, with shorter paths being given the most weight:

$$Katz(s,t) = \sum_{i=1}^{\infty} \beta^i p_i$$

Above, p_i is the number of paths of length i connecting nodes s and t, while β is an input parameter. Here, we have clear separation between the bulk of each distribution. Regardless, for the reasons described above, even the Katz measure is ineffective at predicting links.

Of great interest, however, is the leftward tail of the Katz distribution for positive pairs. This tells us that a small number of new links in the graph appear in structural contexts that appear to be random. It is these links (the statistically unlikely ones) that deserve our interest in the DBLP domain — they represent collaborations between authors from vastly different spheres of the academy.

To verify the ability of the Katz measure to help us identify anomalous links, we conducted a synthetic experiment. We inserted "artificial" links into the data set, randomly selected from the space of all collaborations, and calculated the Katz score for each one. Figure 10 shows where these scores lie in the distribution of actual links (shaded portions of the bars in the histogram represent the synthetic links). As we might expect, the Katz values for the artificial links are grouped in the tail of the overall distribution. However, given the presence of the "interesting" links as described above, its understandable that these actual anomalies could be confused with the apparent ones.

It should be noted that in order to produce non-trivial values of the Katz score (that is, values not equal to zero), the pair of authors in question must be connected in the graph by a minimum path length of four or less. Furthermore, pairs that are not so connected were not considered in our synthetic anomaly experiments. As most randomly selected pairs are indeed not connected by such a short path, it can be assumed that extending our Katz calculations out to paths of length five would shift the mean of the Katz distribution for the synthetically inserted links further out toward the tail of the Katz distribution for actual links.

Finally, we conducted a qualitative examination of a subset of the dataset to test our intuition about the usefulness of the Katz measure to find "interesting" coauthor pairs. From the DBLP data set, we considered the 203 papers from 2003-2004 (representing 450 pairwise collaborations) written by members of the program committee for the workshop on Multirelational Data Mining (MRDM) at KDD 2005, and calculated the Katz score for each pairwise collaboration. The overall distribution of Katz scores for the group is shown in Figure 11. While the bulk of the distribution

of pairs has a score between 0.001 and 0.0001 (note that the histogram is plotted on a log scale), the distribution has a significant tail. It is these collaborations that are often the most interesting, as they are statistically indistinguishable from random pairings in the data. Likewise, pairs with high Katz scores represent links that are not unexpected given the existing structure in the data.

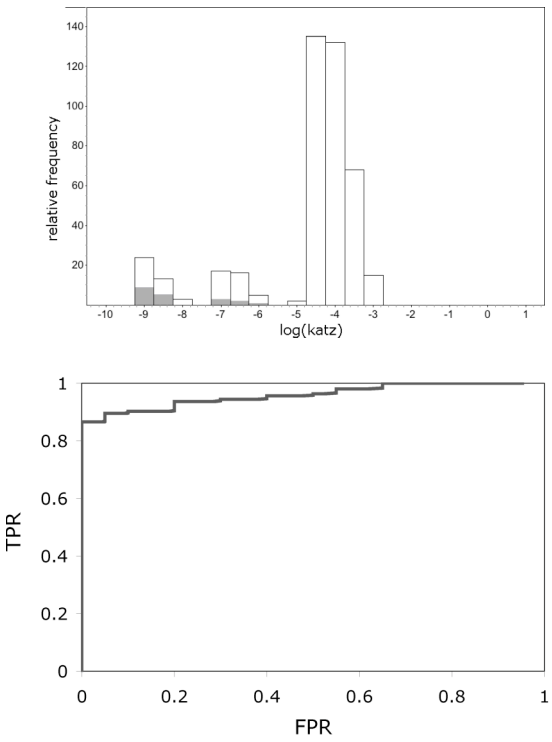


Figure 10. (top) Distribution of log-Katz scores for pairs in DBLP, with artificial links are represented by shaded regions. (bottom) An accompanying ROC curve for classifying anomalies.

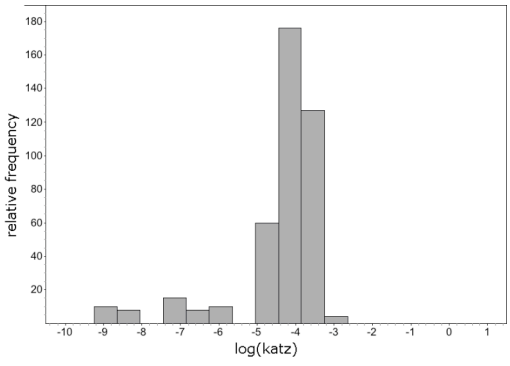


Figure 11. Frequency distribution of log-Katz scores for collaborations involving MRDM program committee members in 2003-2004.

Below we present lists of the ten most "likely" and "unlikely" papers, as ranked by the mean pairwise Katz score for all pairs of authors of the paper. Note that some authors may be omitted, as the data were drawn from the designated pool of "core" authors. In addition, multiple papers written by the same group were disregarded for the sake of interest, and anomalous papers with (unrankable) Katz scores of 0 were excluded.

Top 10 Least Likely Papers of 2003-2004

1. Jean-Francois Boulicaut, Celine Robardet. Constraint-Based Mining Of Formal Concepts In Transactional Data. 2004

2. Thomas Gartner, Stefan Eicker. Einsatz Virtueller Computerpools Im E-Learning. 2003.

3. Jiawei Han, Michael Welge, David Clutter. MAIDS: Mining Alarming Incidents From Data Streams. 2004.

4. Jiawei Han, Michael Garland. Mining Scale-Free Networks Using Geodesic Clustering. 2004.

5. Peter A Flach, Annalisa Appice, Michelangelo Ceci. Redundant Feature Elimination For Multi-Class Problems. 2004.

6. Kristian Kersting, Jorg Fischer. Scaled CGEM: A Fast Accelerated EM. 2003.

7. Jean-Francois Boulicaut, Celine Robardet. Using Classification And Visualization On Pattern Databases For Gene Expression Data Analysis. 2004.

8. Jean-Francois Boulicaut, Bruno Cremilleux. Using Transposition For Pattern Discovery From Microarray Data. 2003.

9. Foster J Provost, Raymond J Mooney, Prem Melville. Active Feature-Value Acquisition For Classifier Induction. 2004.

10. Hiroshi Motoda, Kouzou Ohara, Noboru Babaguchi. Constructive Inductive Learning Based On Meta-Attributes. 2004.

Top 10 Most Likely Papers of 2003-2004

1. Jiawei Han, Jian Pei, Jianyong Wang. CLOSET+: Searching For The Best Strategies For Mining Frequent Closed Itemsets. 2003.

2. Takashi Washio, Hiroshi Motoda. State Of The Art Of Graph-Based Data Mining. 2003.

3. Jiawei Han, Joyce M W Lam, Guozhu Dong, Ke Wang, Jian Pei. Mining Constrained Gradients In Large Databases. 2004.

4. Donato Malerba, Michelangelo Cec, Floriana Esposito. Learning Logic Programs For Layout Analysis Correction. 2003.

5. Hendrik Blockeel, Saso Dzeroski. First Order Random Forests With Complex Aggregates. 2004.

6. Saso Dzeroski, Ljupco Todorovski. Using Domain Specific Knowledge For Automated Modeling. 2003.

7. Donato Malerba, Oronzo Altamura, Teresa Maria Altomare Basile, Nicola Di Mauro, Stefano Ferilli, Michelangelo Ceci, Giovanni Semeraro, Floriana Esposito. Machine Learning Methods For Automatically Processing Historical Documents: From Paper Acquisition To XML Transformation. 2004.

8. Jiawei Han, Ling Feng, Anthony K H Tung, Hongjun Lu. Efficient Mining Of Intertransaction Association Rules. 2003.

9. Raghu Ramakrishnan, Raghav Kaushik, Rajasekar Krishnamurthy, Jeffrey F Naughton. On The Integration Of Structure Indexes And Inverted Lists. 2004.

10. Ashwin Srinivasan, Ross D King. An Empirical Study Of The Use Of Relevance Information In Inductive Logic Programming. 2003.

A cursory inspection of these results confirms our intuitions about the value of the Katz score in determining link likelihood. The papers on the "unlikely" list have author lists similar in character to the Han-Ester collaboration discussed in the previous section, made up of established researchers from different "relational

neighborhoods" within the graph. Omitted from this list is an extremely unlikely paper on "voltage scheduling" written by one of the authors of the paper you are reading now — so unlikely, in fact, that said author does not recall writing it. As it turns out, this lack of recollection is entirely warranted, as there are two computer scientists with the name "David Jensen", and our version of the DBLP database mistakenly combined them.

Conversely, the papers on the second list are written by prolific authors who are multi-connected by shared work and coauthors. Note, for instance, the presence on the list of the two chairs of the MRDM workshop (ranked fifth) as well as teams of authors who have collaborated dozens of times in the literature.

3. RELATED AND FUTURE WORK

The vast majority of work dealing with probabilistic models of link structure addresses the problem of link prediction (e.g., [7], [9], [10]). To our knowledge, though, the simpler problem of anomalous link discovery has been largely ignored. Certainly, ALD could be cast as one of several challenges in structure learning identified by Getoor [3], such as link type prediction (if we consider existential status as a type), link cardinality prediction, or existence and reference uncertainty[4]. Lin and Chalupsky's work on "rarity analysis"[8] seems to be closest in spirit to ALD in terms of motivation and approach, though it focuses on identifying unique paths through the data graph rather than examining the statistical properties of individual links. An obvious extension to ALD would involve combining the two approaches.

While our work on using link prediction models to perform ALD is preliminary, the results so far are encouraging. Further examination and testing on additional data sets (both real and synthetic) is necessary before we gain a complete understanding of how these models work. Future work should also include learning link models that combine a number of statistics (Katz measure, Faloutsos measure, etc.) as well as object attributes to identify interesting links rather than relying on a single measure.

Regardless of method, however, more attention to the anomaly discovery is warranted, because ALD gets at the heart of what defines relational learning as a unique field — the meaning of relationships encoded in the structure of the data.

4. ACKNOWLEDGEMENTS

The authors wish acknowledge the helpful input of Ross M. Fairgrieve and Agustin Schapria.

This research is supported by DARPA and LLNL/DOE LawrenceLivermore National Laboratory and the Department of Energy under contract numbers HR0011-04-1-0013 and W7405-ENG-48. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements either expressed or implied, of DARPA, LLNL/DOE Lawrence Livermore National Laboratory and the Department of Energy, or the U.S. Government.

5. REFERENCES

[1] Faloutsos, C., McCurley, K., and Tomkins, A. Fast Discovery Of Connection Subgraphs. Proc. 10th ACM Conference on Knowledge Discovery and Data Mining, 2004.

[2] Friedman, J. On Bias, Variance, 0/1 Loss, and the Curse of Dimensionality. Data Mining and Knowledge Discovery, p55-77, 1996.

[3] Getoor, L. Link Mining: A New Data Mining Challenge. SIGKDD Explorations, volume 5, issue 1, 2003.

[4] Getoor, L., Friedman, N., Koller, D., and Taskar, B. Learning Probabalistic Models of Link Structure. Journal of Machine Learning Research, 2003.

[5] Jensen, D., Rattigan, M., and Blau, H. Information Awareness: A Prospective Technical Assessment. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.

[6] Lee, W. and Stolfo, S. Data Mining Approaches For Intrusion Detection. Proceedings of the Seventh USENIX Security Symposium (SECURITY '98), 1998.

[7] Liben-Nowell, D. and Kleinberg, J. The Link Prediction Problem For Social Networks. Proc. 12th International Conference on Information and Knowledge Management (CIKM), 2003.

[8] Lin, S. and Chalupsky, H. Unsupervised Link Discovery In Multi-relational Data Via Rarity Analysis. Proceedings of the Third IEEE International Conference on Data Mining, 2003.

[9] Mooney, R., Mellville, P., Tang, L., Shavlik, J., Dutra, I., Page, D., and Costa, V. Relational Data Mining With Inductive Logic Programming For Link Discovery. Proceedings of the National Science Foundation Workshop on Next Generation Data Mining, 2002.

[10] Popescul, A. and Ungar, L. Statistical Relational Learning For Link Prediction. Workshop on Learning Statistical Models from Relational Data at IJCAI 2003.

[11] Taskar, B., Wong, M., Abbeel, P., and Koller, D. Link Prediction in Relational Data. Neural Information Processing Systems Conference (NIPS), 2003